Best Practises for Multilingual Linked Open Data: a Community Effort

Jorge Gracia

Ontology Engineering Group (OEG)
Universidad Politécnica de Madrid (UPM)

jgracia@fi.upm.es

Jose Labra

Web Semantics Oviedo (WESO)
University of Oviedo

labra@uniovi.es

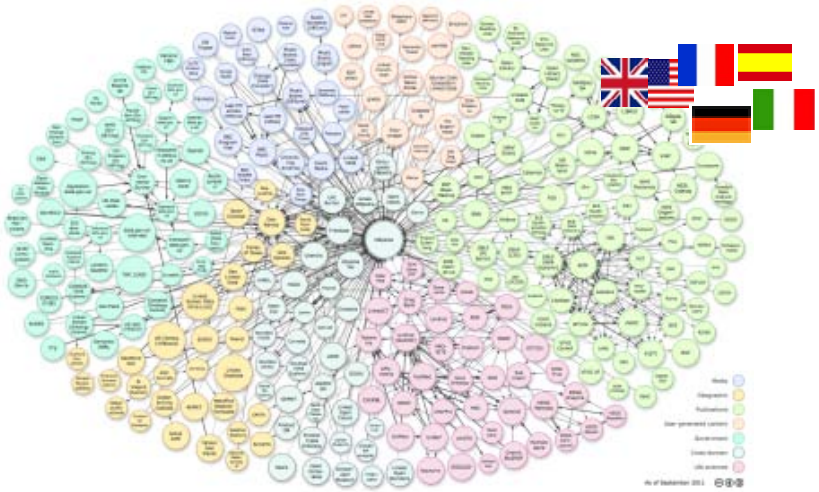
Multilingual Web Workshop
Madrid (Spain)
7-8 May 2014

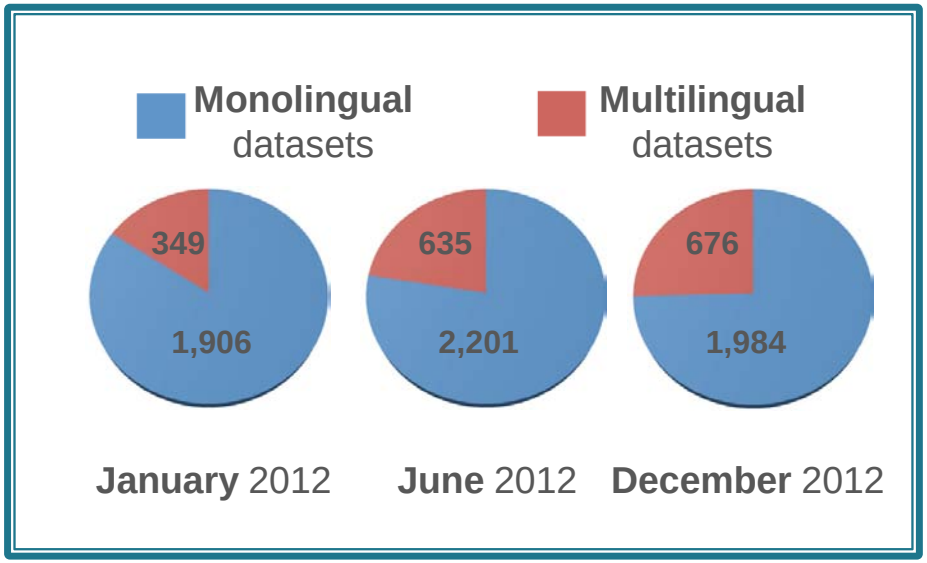
Outline

- Motivation
- The group
- Main goals
- Activities
- Where are we now?

Motivation

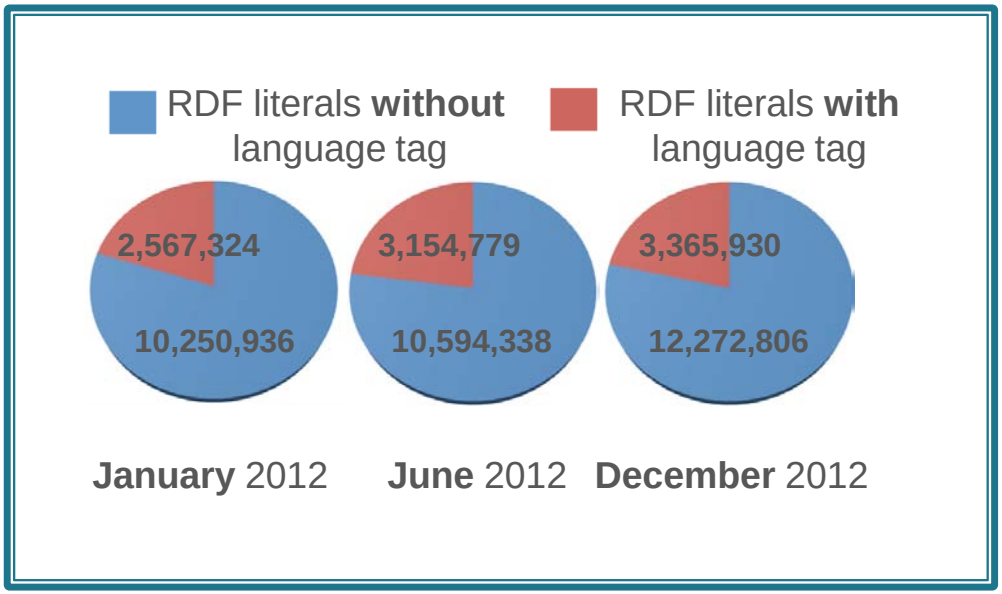
The Web of Data is increasingly multilingual





A. Gómez-Pérez, D. Vila-Suero, E. Montiel-Ponsoda, J. Gracia, and G. Aguado-de Cea, "Guidelines for multilingual linked data," in Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics, ser. WIMS '13. New York, NY, USA: ACM, Jun. 2013.

Standard Semantic Web techniques (e.g. language tagging) are however underused

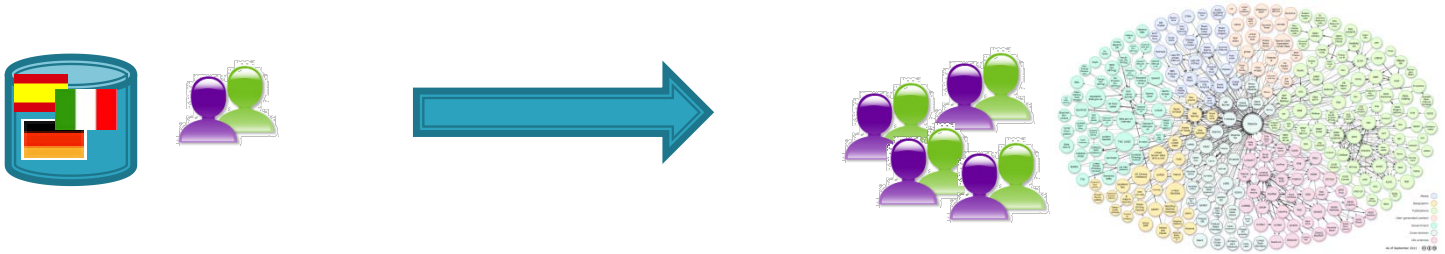


A. Gómez-Pérez, D. Vila-Suero, E. Montiel-Ponsoda, J. Gracia, and G. Aguado-de Cea, "Guidelines for multilingual linked data," in Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics, ser. WIMS '13. New York, NY, USA: ACM, Jun. 2013.

Lots of design decisions have to be made when publishing/consuming/linking multilingual linked data.

For instance…

"How do I publish my data on the Multilingual Web of Data?"



Vocabulary selection

↳ RDF generation

↳ Data Interlinking

↳ Web Publishing

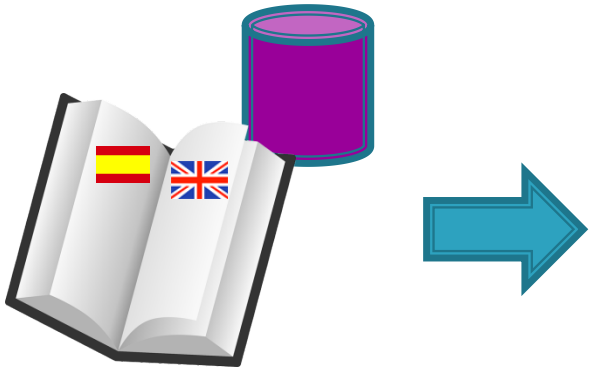
"How do I identify 'things' on the Multilingual Web of Data?"

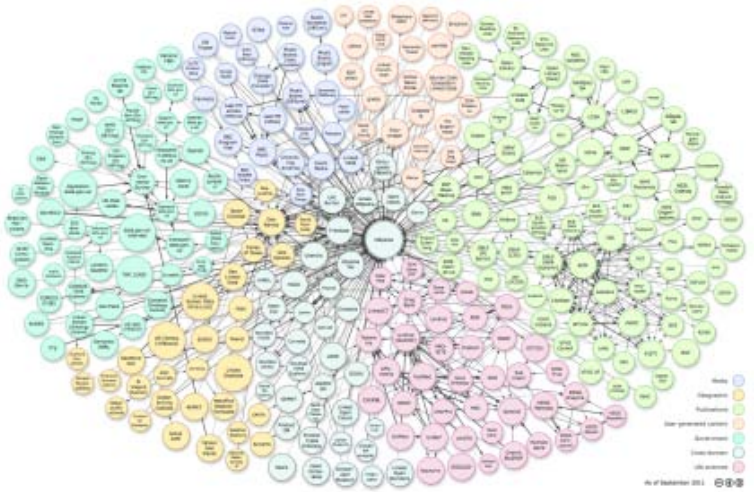
http://example.org/Spain

http://example.org/I23AX45

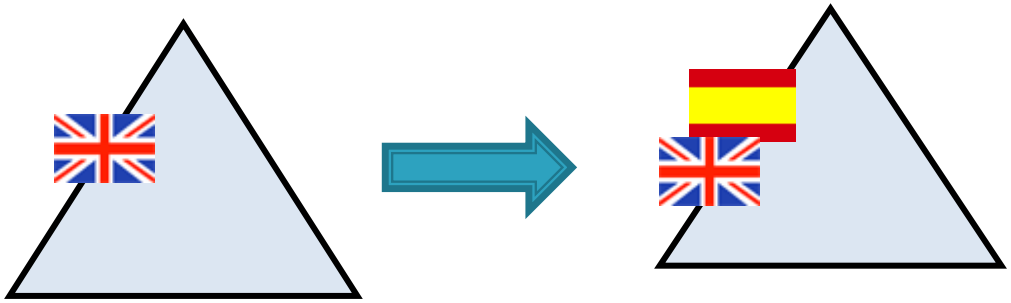
http://example.org/España

"How do I create a Linked Data version of my bilingual dictionary?"





"How do I localise an existing ontology in my own language"



The group

The group

W3C community group on Best Practises for Multilingual Linked (Open) Data

https://www.w3.org/community/bpmlod

Started on June 2013
bi-weekly telcos
3 chairs. Currently:


José Labra


Jorge Gracia


John McCrae

67 members from academia and industry

and many others…

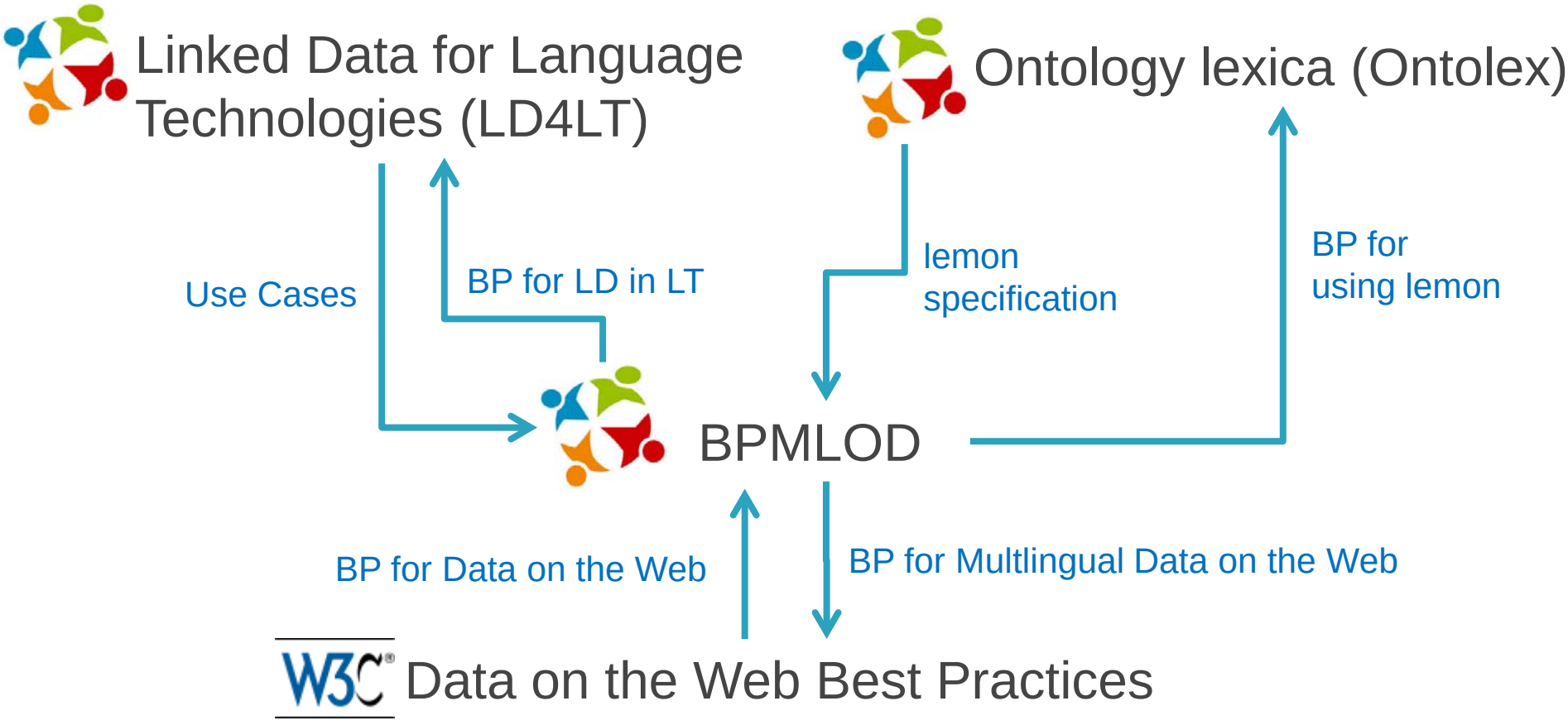
Main goals

Main goals

Crowdsourcing ideas from the community regarding best practices to produce multilingual linked (open) data.

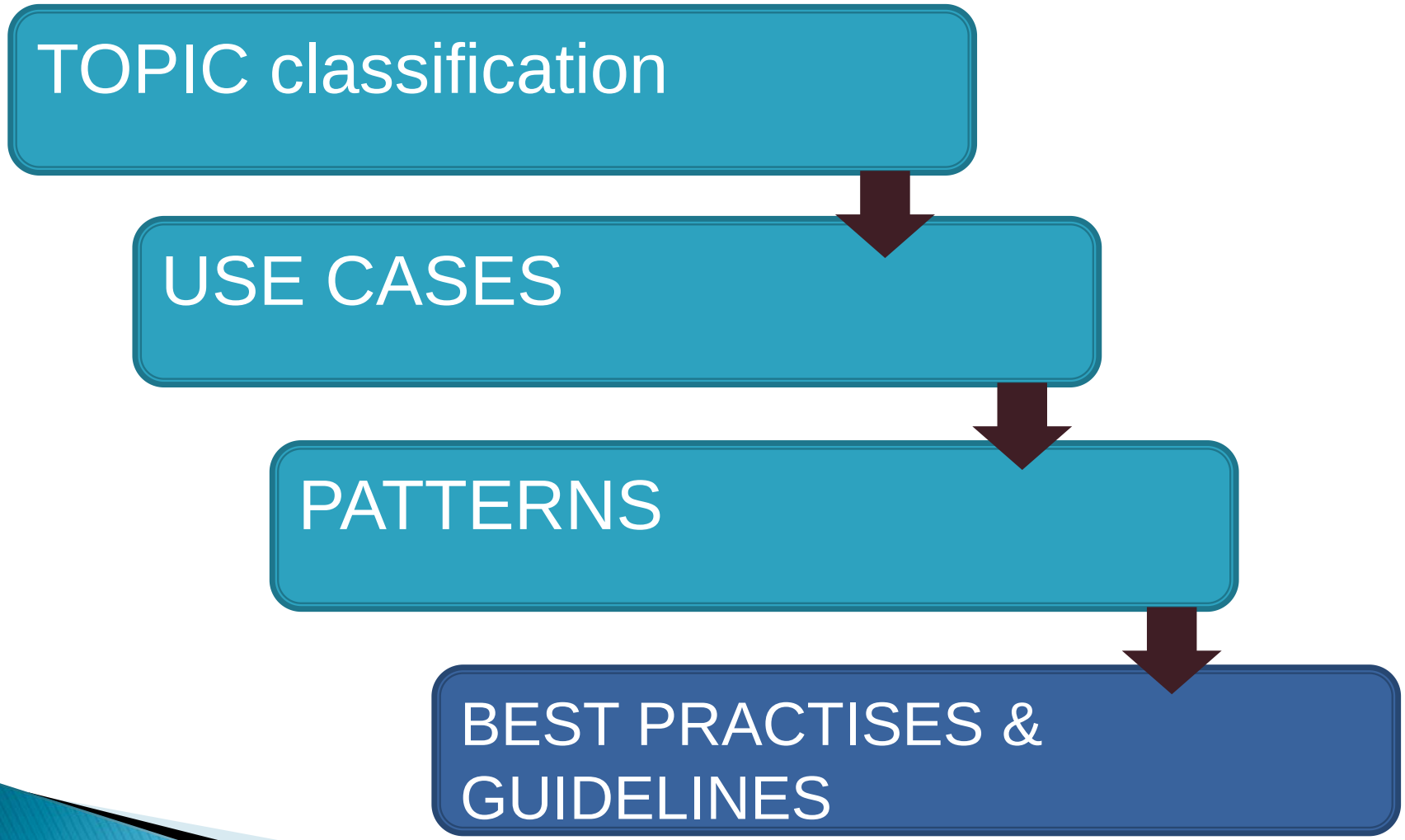
Documenting patterns and best practices for the creation, linking, and use of multilingual linked data.

Relation to other W3C groups



Activities

Activities


TOPIC classification
USE CASES
PATTERNS
BEST PRACTISES & GUIDELINES

TOPIC classification

Topics

- Naming
  - Opaque URIs, Descriptive URIs, IRIs, …
- Textual information
  - Language tags, linguistic information, …
- Linking
  - Interlanguage links, owl:sameAs, …
- Ontologies and vocabularies
  - Mono/multilingual vocabularies, ontology localisation…
- Quality of MLOD
- Tools and examples of MLOD
- Other related aspects
  - licensing, legal aspects, …

https://www.w3.org/community/bpmlod/wiki/Topic_classification

TOPIC classification

USE CASES

Use cases

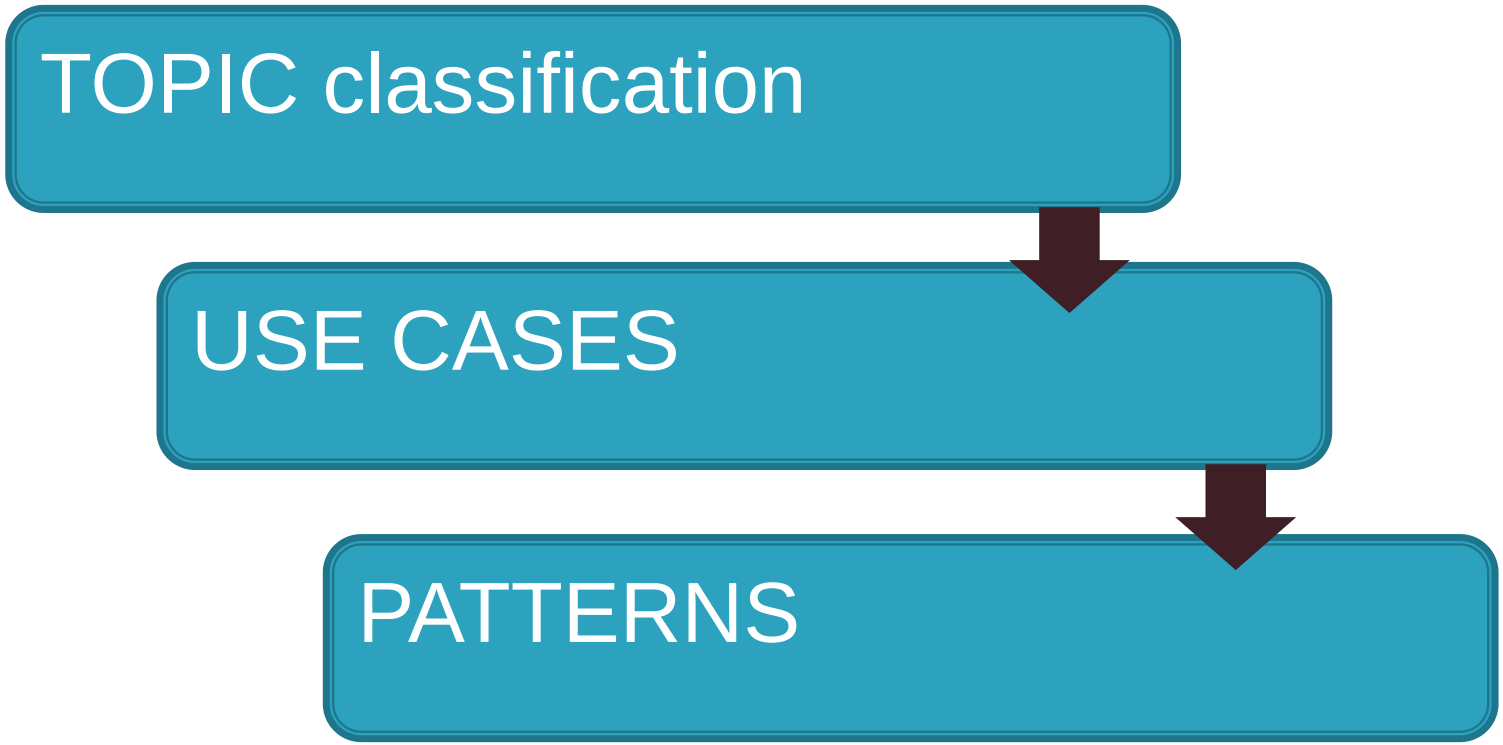
USE CASES

1. Localization workflow [D. Lewis]
2. Lexicalisation of RDF Datasets [E. Montiel, G. Dunshire]
3. Ontology localisation [E. Montiel, L. Aguado, G. Dunsire]
4. Crosslingual linked data matching [J. Gracia]
5. Machine translation [T. Heuss]
6. Application localization [J. McCrae]

CASE STUDIES

1. Translations of multilingual terminologies for libraries [G. Dunsire]

https://www.w3.org/community/bpmlod/wiki/Use_cases_definition


TOPIC classification
USE CASES
PATTERNS

MLOD Patterns

Difficult to establish a boundary between

Patterns vs *Best Practices* vs *Bad smells*

By now: we identify the main practices

Bad/Good may depend on the context/use case

Examples:

- *Patterns for naming and dereferencing*

Patterns for Naming

- Example: URI for Armenia?

Descriptive URIs

http://example.org/Armenia

Human-readable
Good tool support

May be unreadable for non-Latin alphabet users
Difficult to be descriptive enough in some contexts

Independence between concept and language
Maintenance: changes in text don't affect URI

Non Human-readable
Difficult to handle by developers

Readable (for one language)

Security issues (spoofing)
Unreadable for speakers of other languages

Less security issues
Path readable (for one language)

Unreadable for speakers of other languages

tp://en.example.org#Armenia

Practical reasons
Independent development of datasets by language

Where should we put the language tag?
Dialects can become unwieldy
Example: languages & sublanguages

http://example.org/Armenia.en
http://example.org/en/Armenia
http://example.org/Armenia?lang=en

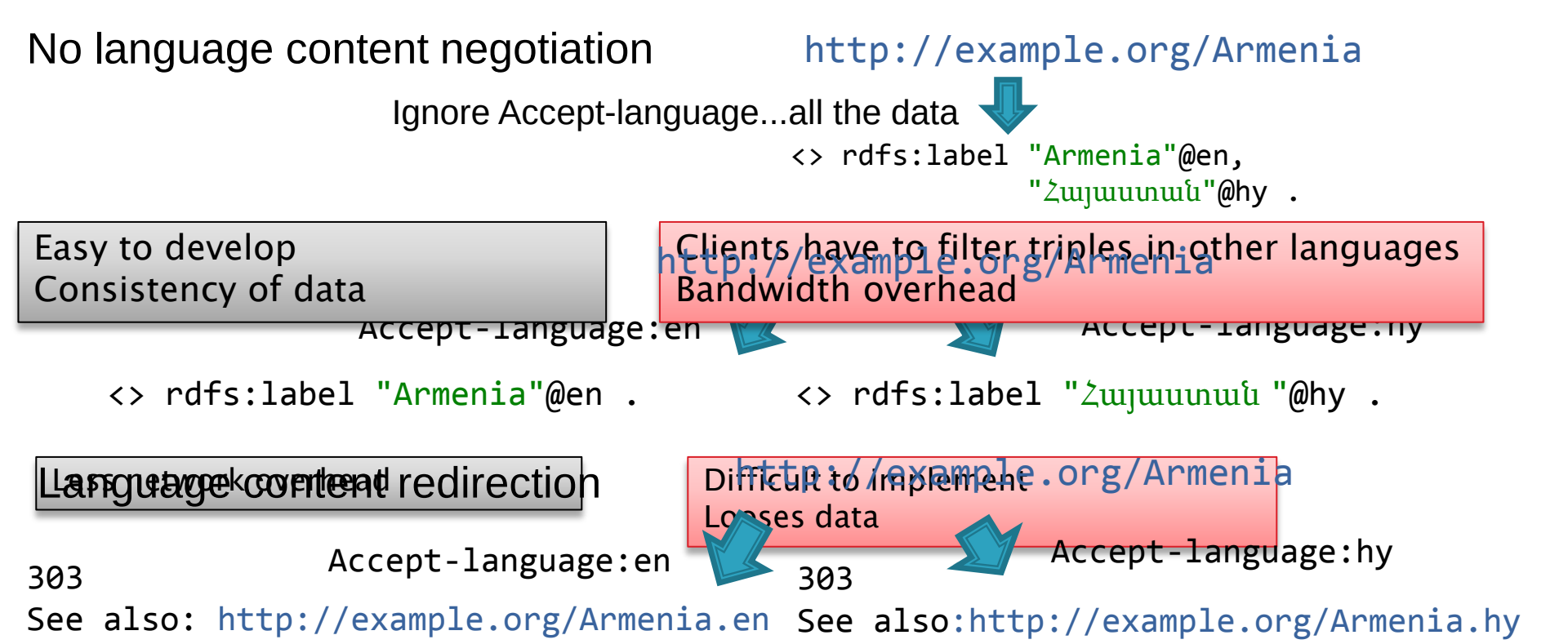
Language in Path

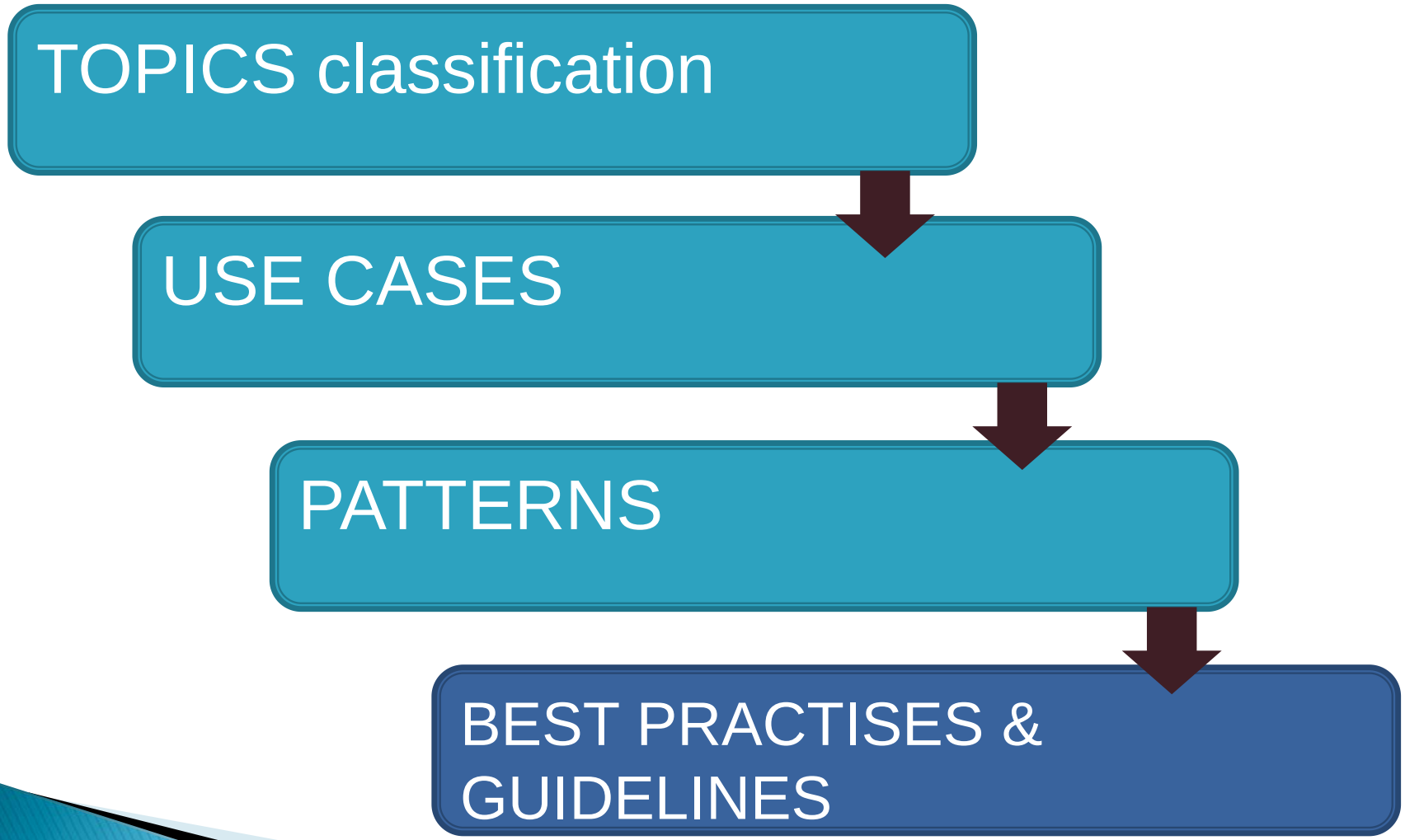
Compatible with content negotiation

Dialects

Patterns for dereferencing

- Which data should I return when accessing a URI?




TOPICS classification
USE CASES
PATTERNS
BEST PRACTISES & GUIDELINES

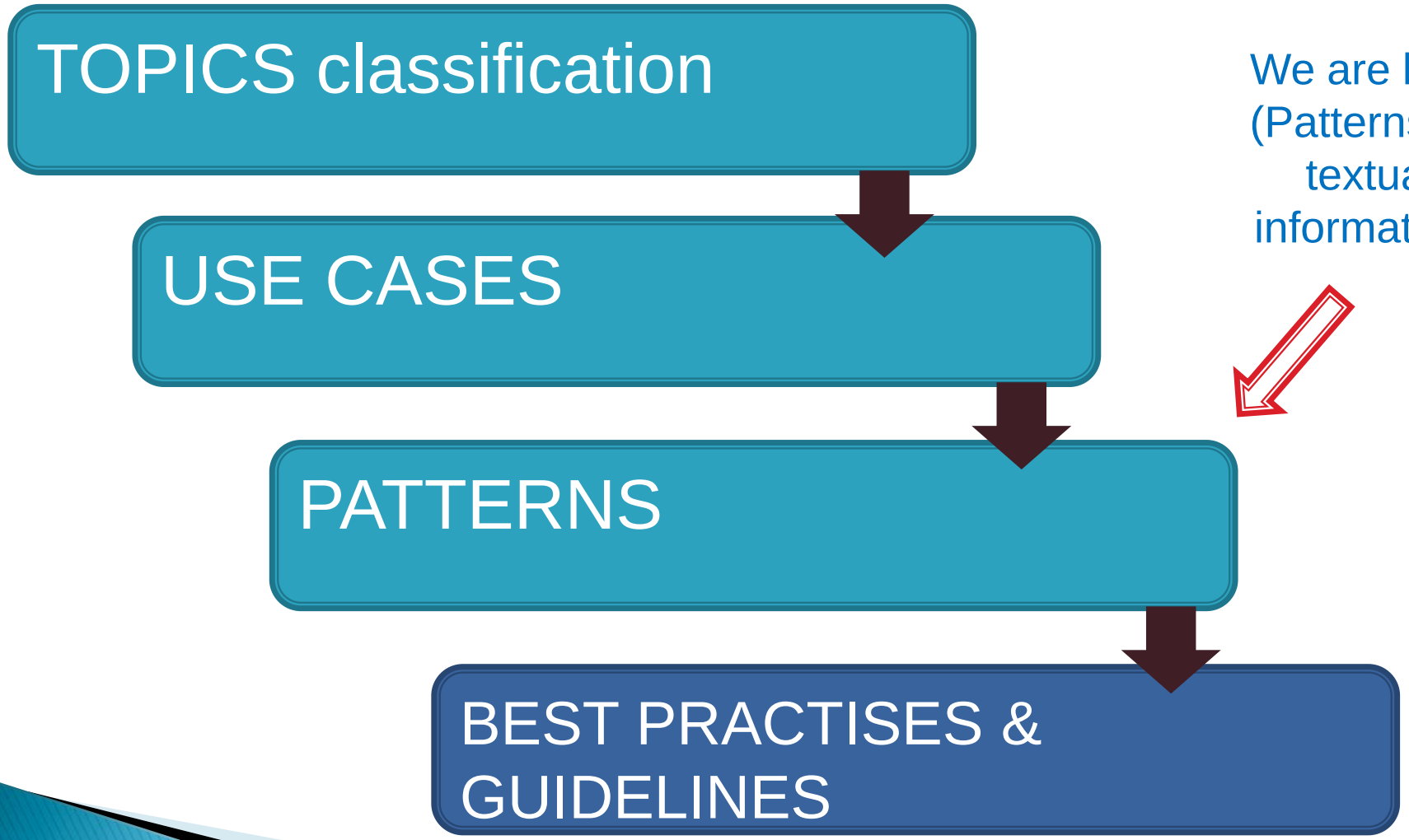
Best practices and guidelines

Some (future) EXAMPLES. Guidelines for:

- Linguistic Linked Data generation
- RDF and Ontology translation
- Multilingual Linked Data generation, publication and exploitation

...

Where are we now?


TOPICS classification
USE CASES
PATTERNS
BEST PRACTISES & GUIDELINES
We are here (Patterns for textual information)

Thanks… and get involved!

Next telco:
Thursday 22nd May
10:00 CEST

https://www.w3.org/community/bpmlod